Establishing normative data for repeated cognitive assessment: A comparison of different statistical methods

Wim Van der Elst, Geert Molenberghs, Martin P. J. Van Boxtel & Jelle Jolles

Behavior Research Methods

e-ISSN 1554-3528

Behav Res
DOI 10.3758/s13428-012-0305-y

Springer

Behav Res
DOI 10.3758/s13428-012-0305-y

Establishing normative data for repeated cognitive assessment: A comparison of different statistical methods

Wim Van der Elst · Geert Molenberghs · Martin P. J. Van Boxtel · Jelle Jolles

Abstract Serial cognitive assessment is conducted to monitor changes in the cognitive abilities of patients over time. At present, mainly the regression-based change and the ANCOVA approaches are used to establish normative data for serial cognitive assessment. These methods are straightforward, but they have some severe drawbacks. For example, they can only consider the data of two measurement occasions. In this article, we propose three alternative normative methods that are not hampered by these problems—that is, multivariate regression, the standard linear mixed model (LMM), and the linear mixed model combined with multiple imputation (LMM with MI) approaches. The multivariate regression method is primarily useful when a small number of repeated measurements are taken at fixed time points. When the data are more unbalanced, the standard LMM and the LMM with MI methods are more appropriate because they allow for a more adequate modeling of the covariance structure. The standard LMM has the advantage that it is easier to conduct and that it does not require a Monte Carlo component. The LMM with MI, on the other hand, has the advantage that it can flexibly deal with missing responses and missing covariate values at the same time. The different normative methods are illustrated on the basis of the data of a large longitudinal study in which a cognitive test (the Stroop Color Word Test) was administered at four measurement occasions (i.e., at baseline and 3, 6, and 12 years later). The results are discussed and suggestions for future research are provided.

Keywords Serial testing · Norms · Practice effects · Longitudinal data · Linear mixed model · Multiple imputation · Stroop Color Word Test

W. Van der Elst (✉) · G. Molenberghs
Interuniversity Institute for Biostatistics and Statistical Bioinformatics (CenStat), Universiteit Hasselt, Martelarenlaan 42, 3500 Hasselt, Belgium
e-mail: Wim.Vanderelst@gmail.com

G. Molenberghs
Katholieke Universiteit Leuven (Belgium), Leuven, Belgium

M. P. J. Van Boxtel
School for Mental Health and Neuroscience (MHeNS), and European Graduate School of Neuroscience (EURON), Maastricht University, Maastricht, The Netherlands

J. Jolles
LEARN! Research Institute and Faculty of Psychology & Education, VU Universiteit Amsterdam, Amsterdam, The Netherlands

Published online: 24 January 2013

Cognition is an umbrella term that refers to various higher-order behavioral abilities, such as memory, attention, and executive functions (Lezak, Howieson, & Loring, 2004). These higher-order behavioral abilities are latent variables that cannot be directly observed. Instead, they have to be inferred from proxy measures (Mitrushina, Boone, Razani, & D'Elia, 2005). For example, a person's verbal memory cannot be directly observed; what can be observed is the person's ability to recall verbal material that is presented in a specific standardized test setting.

Cognitive assessment is widely used in medical settings and in the behavioral sciences—for example, in clinical psychology, educational practice, and rehabilitation settings (Lezak et al., 2004; Pasquier, 1999). In diagnostic settings, the "raw" score of a person on a cognitive test (e.g., the number of items that were recalled in a memory test) is usually not of direct interest. The reason for this is that the raw scores on cognitive tests are strongly affected by demographic variables (such as age and educational level; Mitrushina et al., 2005; Strauss, Sherman, & Spreen, 2006; Van der Elst, 2006). For example, the same raw test score may be indicative of a severe memory problem in a 50-year-old person, while it is within the normal limits of test performance for an 80-year-old person (Van der Elst, Van

Boxtel, Van Breukelen, & Jolles, 2005). Clinicians therefore use relative measures (rather than raw test scores) to evaluate a patient's test performance (e.g., what is the percentage of demographically matched "cognitively healthy" peers who obtain a test score that is equal to or worse than the test score of this patient?). So-called normative data are used to convert raw test scores into demographically corrected relative measures (Mitrushina et al., 2005; Van der Elst, 2006).

In many diagnostic situations, the same cognitive test (or a parallel test version) is repeatedly administered to the same person. For example, a clinician may need to determine whether a patient with mild cognitive impairment has experienced cognitive decline since his or her last evaluation, or a clinician may need to evaluate whether a stroke patient has benefited from taking part in a rehabilitation program. Ideally, the observed changes in the test scores at subsequent measurement occasions would be directly interpretable in terms of true changes in the latent cognitive trait of interest. This is, however, generally *not* the case (Calamia, Markon, & Tranel, 2012). The main reason for this is that practice effects occur in serial testing situations. Practice effects refer to a variety of factors—such as procedural learning, memory for specific items, and increased comfort with formal testing situations (McCaffrey, Duff, & Westervelt, 2000)—that result in systematic improvements in test scores at retesting occasions, even though there was no true change in the latent trait that is measured by the cognitive test (Bartels, Wegrzyn, Wiedl, Ackermann, & Ehrenreich, 2010; Calamia et al., 2012; Dikmen, Heaton, Grant, & Temkin, 1999; Temkin, Heaton, Grant, & Dikmen, 1999; Van der Elst, Van Breukelen, Van Boxtel, & Jolles, 2008). Practice effects are especially pronounced when the test–retest intervals are short (e.g., Theisen, Rapport, Axelrod, & Brines, 1998), but they also occur in studies with test–retest intervals of several years (Rönnlund & Nilsson, 2006; Salthouse, Schroeder, & Ferrer, 2004). In the latter case, the changes in the test scores over time reflect the combined influences of practice effects and true changes in the latent cognitive abilities (Van der Elst et al., 2008). Furthermore, the extent to which practice effects occur is affected by person characteristics such as the age and the educational level of a tested person (Mitrushina & Satz, 1991; Rapport, Brines, Axelrod, & Theisen, 1997; Stuss, Stethem, & Poirier, 1987; Van der Elst et al., 2008).

Failure to take practice effects into account may invalidate the conclusions that are drawn from a serial cognitive assessment (Calamia et al., 2012; Van der Elst et al., 2008). For example, practice effects may mask the cognitive decline in a patient with early dementia, or practice effects may lead to the incorrect conclusion that a stroke patient has benefitted from a rehabilitation program. Normative data for serial cognitive assessment should thus take the testing history and the demographic characteristics of a patient into account, but it is not clear which statistical method is optimal for achieving this aim (Heaton et al., 2001; Temkin et al., 1999; Van der Elst et al., 2008).

Existing normative methods

In nonserial (i.e., single-measurement) cognitive testing situations, normative data are established on the basis of classical univariate statistical methods. For example, an often-used procedure is the regression-based normative approach (Testa, Winicki, Pearlson, Gordon, & Schretlen, 2009; Van Breukelen & Vlaeyen, 2005; Van der Elst et al., in press; Van der Elst et al. 2006a, 2006b, 2006c, 2006d). In this method, a classical multiple linear regression model is fitted to the data of a large sample of cognitively healthy people who were administered the cognitive test of interest (the normative sample). The multiple linear regression model assumes that $\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon_i}$ where $\boldsymbol{Y}_i$ is the vector of the responses, $\boldsymbol{X}_i$ is the design matrix (which typically includes age, gender, and educational level in normative studies), $\boldsymbol{\beta}$ is the vector of regression parameters, and $\boldsymbol{\varepsilon}_i$ is the vector of the residual components (for details on this model, see, e.g., Kutner, Nachtsheim, Neter, & Li, 2005).

On the basis of the established regression model, the test performance of a future patient j can be evaluated. This requires three steps. First, the expected test score of patient j is computed (i.e., $\widehat{Y}_j = \boldsymbol{X}_j\widehat{\boldsymbol{\beta}}$). This score reflects the expected test score for a cognitively healthy person who has the same demographic background as the tested patient. Second, the difference between the patient's observed and expected test scores is computed (i.e., $e_j = Y_j - \widehat{Y}_j$) and standardized (i.e., $Z_j = e_j / SD(e)$). The $SD(e)$ is the SD of the residuals in the normative sample (which is approximately equal to the positive square root of the residual mean squares). Third, the standardized residual of the patient is converted into a percentile value as based on the distribution of the standardized residuals in the normative sample. A percentile value below 5 is often considered as being indicative of a cognitive problem (because 95 % of the "cognitively healthy" people perform better).

An important assumption of the classical linear regression model is that $\sigma^2\{\boldsymbol{\varepsilon}\} = \sigma^2\boldsymbol{I}$ (with $\boldsymbol{I}$ = an $n \times n$ identity matrix). Thus, it is assumed that the residuals (or equivalently, the responses) are uncorrelated. This assumption is not realistic in serial cognitive-testing situations, because the cognitive test scores at subsequent measurement occasions tend to be highly correlated within individuals (Dikmen et al., 1999; Lezak et al., 2004; Temkin et al., 1999; Van der Elst et al., 2008). One possible solution for dealing with this problem is to summarize the vector of the repeated

measurements into change scores (change = endpoint score − baseline score) and, subsequently, regress these responses on the demographic covariates of interest in the normative sample (the regression-based change approach). Alternatively, the dependence issue can be solved by fitting a model in which the endpoint scores are regressed on the baseline scores and the demographic covariates in the normative sample (the ANCOVA approach).

Motivating example

To illustrate the problems with the existing normative methods and to exemplify the newly proposed methods (see below), data from the Maastricht Aging Study (MAAS) are used. The MAAS is a longitudinal research project into the determinants of cognitive aging (Jolles, Houx, Van Boxtel, & Ponds, 1995). The MAAS baseline measurement took place between 1993 and 1996, and three follow-up measurements were conducted (3, 6, and 12 years after baseline). All participants were thoroughly screened for medical pathology that could interfere with normal cognition, such as dementia or cerebrovascular disease.

The MAAS participants were administered an extensive battery of cognitive and medical tests. In the present article, we will focus on the data of the Stroop Color Word Test (SCWT; Stroop, 1935). The SCWT is a well-known cognitive paradigm that is used to assess inhibition and other components of executive functioning (Lezak et al., 2004; Moering, Schinka, Mortimer, & Graves, 2004). The test consists of three subtasks. The first subtask shows color words in random order (red, blue, yellow, green) that are printed in black ink. The second subtask displays solid color patches in one of these four basic colors. The third subtask contains color words that are printed in an incongruous ink color (e.g., the word "red" printed in yellow ink). The participants are instructed to read the words, name the colors, and name the ink color of the printed words as quickly and as accurately as possible in the three subsequent subtasks. The SCWT outcome variable of interest is the difference between the time that is needed to complete subtask three and the average time that is needed to complete the first two subtasks [i.e., SCWT score = time in seconds needed for subtask 3 − (time in seconds needed for subtasks 1 + 2) / 2]. Higher SCWT scores are thus indicative of worse test performance.

In the MAAS, the SCWT was administered to $N = 887$, $N = 696$, $N = 614$, and $N = 454$ participants at the subsequent measurement occasions. Missingness in the responses was thus substantial. Basic demographic data for the sample at baseline and at the three follow-up measurement occasions are provided in Table 1. Level of Education (LE) was categorized into three levels using a classification scheme that is often used in the Netherlands (De Bie, 1987), with low = at most primary education, average = at most junior vocational training, and high = senior vocational or academic training. More details regarding the SCWT and the sample frame, participant recruitment, stratification criteria, and other aspects of the MAAS can be found elsewhere (Jolles et al., 1995; Van der Elst, 2006).

Limitations of the existing normative methods

Suppose that the regression-based change method or the ANCOVA approach were used to establish normative data for serial SCWT administration (as based on the MAAS data). This would have two major drawbacks.

First, the ANCOVA and the regression-based change approaches cannot handle missing data appropriately. Both methods simply discard incomplete cases from the analyses, but a complete case analysis is unbiased only when the responses are Missing Completely At Random (MCAR; Little & Rubin, 1987; Rubin, 1976), and even then it is usually inefficient (Verbeke & Molenberghs, 2000). MCAR means that the probability of an observation being missing is independent of the observed or unobserved responses. The MCAR assumption is not realistic in most serial testing settings. For example, the probability that a participant drops out of the MAAS is strongly affected by his or her baseline cognitive test score (Van Beijsterveldt et al., 2002), and thus the MCAR assumption is not valid.

Second, the regression-based change and the ANCOVA methods can only use the data for a maximum of two measurement occasions. In the MAAS, the SCWT was administered four times. The application of the regression-based change or the ANCOVA approach would thus result in a substantial loss of information and, consequently, a lowered precision of the parameter estimates and a loss of power (Verbeke & Molenberghs, 2000). Note that it might be argued that the endpoint score could be regressed on the test scores of multiple earlier testing occasions in the ANCOVA method (rather than on a single one), but this is generally not the case because the test scores at subsequent measurement occasions are highly correlated and, thus, collinearity issues would arise.

Alternative normative methods: Multivariate regression, the standard linear mixed model, and the linear mixed model combined with multiple imputation

As was noted in the previous sections, normative data for serial cognitive assessment should take the testing history and the demographic characteristics of a patient into

Table 1 Demographic characteristics of the participants who were administered the Stroop Color Word Test at the different measurement occasions

Measurement moment	Age group (age at baseline)	*N*	Age (at baseline)		Level of education		Female:Male	
			M	*SD*	High	Average	Low	ratio
Baseline	<60 years	286	54.71	2.94	129	102	55	137:149
	>60 and ≤70 years	319	64.71	3.23	161	118	40	153:166
	>70 years	282	74.72	3.23	145	92	45	141:141
	Total	887	64.67	8.59	435	312	140	431:456
First follow-up	<60 years	236	54.67	2.93	96	91	49	105:131
	>60 and ≤70 years	266	64.67	3.16	128	106	32	121:145
	> 70 years	194	74.31	3.09	101	61	32	102:92
	Total	696	63.97	8.29	325	258	113	328:368
Second follow-up	<60 years	229	54.57	2.96	96	85	48	111:118
	>60 and ≤70 years	240	64.57	3.15	116	94	30	111:129
	>70 years	145	73.88	2.81	76	43	26	81:64
	Total	614	63.04	8.03	288	222	104	303:311
Third follow-up	<60 years	210	54.63	3.02	90	79	41	105:105
	>60 and ≤70 years	179	64.22	3.02	86	71	22	89:90
	>70 years	65	73.68	2.7	31	22	12	43:22
	Total	454	61.14	7.4	207	172	75	237:217

Note. The first, second, and third follow-up measurements were conducted 3, 6, and 12 years after baseline, respectively.

account, but it is not clear which statistical method is optimal to achieve this aim. The existing methods (i.e., the regression-based change and the ANCOVA methods) are fundamentally flawed. Applying these methods to the SCWT data (from the MAAS) would lead to a substantial loss of information and biased results. What we need are (1) methods that can deal with two or more correlated responses (within individuals) and (2) methods that can handle missing data appropriately.

On the basis of these criteria, the use of the multivariate regression model, standard linear mixed model, and linear mixed model with multiple imputation approaches are proposed in the present article. These methods are described in the next sections.

The multivariate regression model

The multivariate regression model assumes that $\boldsymbol{Y}_i = \boldsymbol{X}_i \boldsymbol{\beta} + \varepsilon_i$ with $\boldsymbol{Y}_i$ = the vector of the repeated measurements for subject i ($1 \leq i \leq N$, with N = the number of subjects), $\boldsymbol{X}_i$ = the design matrix, $\boldsymbol{\beta}$ = the vector of the regression parameters, and ε_i = the vector of the error components. It is assumed that $\varepsilon \sim N(\mathbf{0}, \sum)$, with $\mathbf{0}$ = a zero matrix and $\sum$ = a general (unstructured) variance—covariance matrix of the residuals (for details on the model, see Johnson & Wichern, 2007).

In contrast to the classical (or univariate) linear regression model, the multivariate regression model can handle vectors of *repeated* observations for individuals. The parameter estimates in the multivariate regression model are based on likelihood methods, which allow for using all available outcomes in the calculations (Molenberghs & Kenward, 2007). Moreover, the use of likelihood-based methods has the advantage that inferences can be based on the observed likelihood given a model that does not include a distribution for the missing data mechanism (Little & Zhang, 2011; Molenberghs & Verbeke, 2005; Verbeke & Molenberghs, 2000). These so-called *ignorable* analyses require that the missingness mechanism is Missing At Random (MAR; i.e., the probability of an observation being missing is independent of the unobserved outcomes conditional on the observed data) or MCAR (as defined above) when likelihood or Bayesian inferences are chosen, though this assumption can be relaxed in the context of normative analyses (see the Discussion section). Note that the parameter estimates in a multivariate regression model can also be based on ordinary least squares methods (rather than on likelihood-based methods), but this situation will not be considered here because it largely suffers from the same drawbacks as the regression-based change and the ANCOVA methods.

The standard linear mixed model

The random-effects approach toward extending the classical linear regression model to a longitudinal setting is based on the assumption that the responses of a participant can be

appropriately modeled on the basis of a linear regression model in which subject-specific regression coefficients are used (Verbeke & Molenberghs, 2000). In particular, the standard Linear Mixed Model (LMM) assumes that $\boldsymbol{Y}_i = \boldsymbol{X}_i \boldsymbol{\beta} + \boldsymbol{Z}_i \boldsymbol{b}_i + \boldsymbol{\varepsilon}_i$, with $\boldsymbol{Y}_i$ = the vector of the repeated measurements for subject i ($1 \leq i \leq N$, with N = the number of subjects), $\boldsymbol{X}_i$ = the design matrix for the fixed effects (i.e., the population-averaged parameters), $\boldsymbol{\beta}$ = the vector of regression coefficients, $\boldsymbol{Z}_i$ = the design matrix for the subject-specific effects (capturing how individuals deviate from the population average, where the population is understood as any subject with the same fixed-effect design), $\boldsymbol{b}_i$ = the vector of the random effects, and $\boldsymbol{\varepsilon}_i$ = the vector of the residual components. Because the participants in a study are a random sample of a larger population, it is natural to assume that $\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{D})$, where $\mathbf{0}$ is a zero matrix and $\boldsymbol{D}$ is a general variance—covariance matrix. It is furthermore assumed that $\boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \sum_i)$, where $\sum_i$ is a variance—covariance matrix (which was chosen to be equal to $\sigma^2 \boldsymbol{I}_{n_i}$ in the present study, with $\boldsymbol{I}_{n_i}$ = an identity matrix of dimension n_i). The random components $\boldsymbol{b}_1,\ldots,\boldsymbol{b}_n$, $\boldsymbol{\varepsilon}_1,\ldots,\boldsymbol{\varepsilon}_N$ are assumed to be independent. For more details on the estimation and inference for the marginal model and the variance components in a LMM, the reader is referred to Verbeke and Molenberghs (2000).

As compared with the multivariate regression model, the standard LMM has the additional advantage that both fixed and random effects can be included in the model. Random effects are not of substantive interest in normative studies (the focus is on the marginal evolutions—i.e., on the fixed effects), but it is nevertheless useful to model the covariance structure adequately, because this generally leads to more efficient inferences for the fixed effects (i.e., smaller standard errors; Verbeke & Molenberghs, 2000). This is particularly important in the case of unbalanced data—that is, when different subjects provide different numbers of outcome values (either by design or because of missingness in the data). As was also the case for the multivariate regression model, the standard LMM has the advantage that it can take the uncertainty of dealing with missing values into account in the analyses (Verbeke & Molenberghs, 2000).

The linear mixed model combined with multiple imputation

In the LMM combined with Multiple Imputation (LMM with MI) approach, the MI algorithm is first applied to fill in the missing observations in the data set. The key idea of MI is to replace each missing value with M plausible values (Rubin, 1996). Each value can be seen as a Bayesian drawn from the conditional distribution of the unobserved responses given the observed ones (Beunckens, Molenberghs, & Kenward, 2005; Little & Rubin, 1987).

To fix ideas on this method, let us consider a problem where we have two unknown parameters γ_1 and γ_2 and a data set y. In a Bayesian context, these have a joint posterior distribution $\mathrm{f}(\gamma_1, \gamma_2|\mathrm{y})$. Suppose that parameter γ_2 is of interest and that γ_1 is a nuisance parameter (i.e., a parameter that is not of substantive interest but that has to be accounted for in the analysis). The posterior can be partitioned as $\mathrm{f}(\gamma_1, \gamma_2|\mathrm{y}) = \mathrm{f}(\gamma_1|\mathrm{y})\mathrm{f}(\gamma_2|\gamma_1, \mathrm{y})$, so it follows that the marginal posterior for γ_2 can be expressed as $\mathrm{f}(\gamma_2|\mathrm{y}) = \mathrm{E}_{\gamma_1}(\mathrm{f}(\gamma_2|\gamma_1, \mathrm{y}))$, with E_{γ_1} = the expectation over the distribution of γ_1. The posterior mean and variance for γ_2 equal $\mathrm{E}(\gamma_2|\mathrm{y}) = \mathrm{E}_{\gamma 1}\left(\mathrm{E}_{\gamma_2}(\gamma_2|\gamma_1, \mathrm{y})\right)$ and $var(\gamma_2|y) = E_{\gamma 1}\left(var_{\gamma 2}(\gamma_2|\gamma_1, y)\right) + var_{\gamma_1}\left(E_{\gamma_2}(\gamma_2|\gamma_1, y)\right)$, with $\mathrm{var}_{\gamma 2}$ = the variance computed over the distribution of γ_2. These quantities can be approximated by way of empirical moments. Let γ_1^m be draws from the marginal posterior distribution of γ_1 for $m = 1, \ldots, M$. It holds that $E(\gamma_2|y) \cong \frac{1}{M}\sum_{\mathrm{m}=1}^{\mathrm{M}} \left(\mathrm{E}_{\gamma 2}(\gamma_2|\gamma_1^m, \mathrm{y})\right)$. Defining the right hand side of the previous equation as $\tilde{\gamma}_2$, it furthermore holds that $var(\gamma_2|\mathrm{y}) = \frac{1}{M}\sum_{\mathrm{m}=1}^{\mathrm{M}} \mathrm{var}_{\gamma_2}(\gamma_2|\gamma_1^{\mathrm{m}}, \mathrm{y}) + \frac{1}{\mathrm{M}-1}\sum_{\mathrm{m}=1}^{\mathrm{M}} \left(\mathrm{E}_{\gamma 2}(\gamma_2|\gamma_1^{\mathrm{m}}, \mathrm{y}) - \tilde{\gamma}_2\right)^2$. In the MI procedure, these formulas are generalized for vector-valued parameters, and γ_2 is used to represent the substantive model where γ_1 is used to represent the missing data. In sufficiently large samples, the conditional posterior moments for γ_2 can be approximated by maximum likelihood estimators from the completed data set. The MI estimates of the parameters of interest and their variance approximate the first two moments of the posterior distribution in a fully Bayesian analysis (Kenward & Carpenter, 2007).

After imputing the missing values M times using Bayesian draws, a LMM analysis is conducted on each of the completed data sets (or any analysis of interest in a context other that the one considered here), and the different inferences are subsequently combined into a single one (for more details on MI, see chap. 9 of Molenberghs & Kenward, 2007).

As was also the case with the multivariate regression model and the standard LMM, the LMM with MI takes the uncertainty of dealing with missing values into account in the analyses (Rubin, 1996; Verbeke & Molenberghs, 2000). This is to be contrasted with so-called simple imputation methods, in which each missing response is substituted by a single value (Molenberghs & Verbeke, 2005). The standard LMM and the LMM with MI methods are largely equivalent (provided that the imputation model includes all relationships that will be considered in the analyses and the inference tasks; Molenberghs & Kenward, 2007), but the MI method allows for some additional flexibility in dealing with complex data sets. That is, MI can be used to deal with missing covariate values and missing responses at the same time (Molenberghs & Kenward, 2007; Molenberghs & Verbeke, 2005).

Application to the motivating example

In this section, we will illustrate the use of the multivariate regression, standard LMM, and LMM with MI methods to establish norms for serial SCWT administration (as based on the MAAS data described earlier). All analyses were conducted with R 2.14.0 for OS X and SAS v9.2 for Windows. An α-level of .05 was used.

The multivariate regression model

The initial multivariate regression model included the vector of the log(SCWT) scores as the outcome and age, age^2, gender, LE low, LE high, time, and time^2 as the covariates. The SCWT score was log-transformed because preliminary analyses showed that the residuals were positively skewed. Age was centered (age = calendar age in years − 65) prior to the computation of the quadratic age effect (to avoid multicollinearity; Kutner et al., 2005). Gender was coded as 1 = male and 0 = female. The three levels of education (LEs) were coded with two dummies—that is, LE low, 1 = at most primary education and 0 = otherwise; and LE high, 1 = senior vocational or academic training and 0 = otherwise. Time was dummy coded using three dummies and baseline measurement as the reference category. In addition to the main effects, the age × time, age × time^2, age^2 × time, age^2 × time^2, LE low × time, LE high × time, LE low × time^2, and LE high × time^2 interaction terms were included in the mean structure of the initial model. This was done because previous studies have suggested that older age and lower LEs are associated with a more pronounced cognitive decline over time (see, e.g., Salthouse, 1996; Schmand, Smit, Geerlings, & Lindeboom, 1997; Stern, 2003; Van der Elst et al., 2006d).

The initial model had a $-2\,l$ value that equaled 648.2 (see model 1 in Table 2). To obtain the most parsimonious model, it was first evaluated whether the mean structure of the initial model could be simplified by removing interactions and main effect terms. Likelihood ratio tests suggested that the model fit did not significantly deteriorate when the LE × time and the age^2 × time interaction terms were removed from the model (all ps > .05; see models 2 and 3 in Table 2). Next, it was evaluated whether the mean structure could be simplified by assuming linear and quadratic effects of time (instead of using dummies to model the effects of time). In these models, time was centered (time = time since baseline in years − 5.25) prior to the computation of the quadratic terms (to avoid multicollinearity). The likelihood ratio tests suggested that the models in which linear (model 4) and quadratic (model 5) time effects were assumed both adequately fitted the data (using model 3 as the comparison model). The linear, rather than the quadratic, model was retained because it is more parsimonious.

Next, it was evaluated whether age group-, gender-, or LE-specific covariance structures were needed. Age group was constructed on the basis of a median split of the continuous variable age (i.e., younger = ≤65 years at baseline, older = >65 years at baseline). Smoothed (loess) average trends of the squared ordinary least square residuals $\left\{\sigma^2(t) = E\left[Y(t) - \widehat{\mu}(t)^2\right]\right\}$ were plotted for the different subgroups (using age, age^2, gender, LE low, LE high, time, and age × time as covariates in the model). As is shown in Fig. 1, the residual variances for older people tended to be higher, as compared with the residual variances for younger people, at most of the measurement moments. A separate residual variance–covariance matrix $\sum$ was thus fitted for older and younger people (model 6), and this model indeed had a significantly better fit to the data than did model 4 (see Table 2). The variance functions for males and females and for people with a low, average, and high LE were similar (figures not shown), suggesting that gender- and LE-specific covariance structures are not needed.

The most parsimonious multivariate regression model that still adequately fitted the data was thus model 6. The parameter estimates for this model are provided in Table 3a. As is shown, males and lower educated participants had significantly higher log(SCWT) scores at all measurement moments. There was a significant time × age interaction term, which suggested that the increase in the log(SCWT) scores over time was more pronounced for people who were older at baseline. The interaction is graphically depicted in

Table 2 Likelihood ratio tests to evaluate the fit of a series of nested multivariate regression models

Model	Model structure	Number of pars.	$-2\,l$	Ref. Model	$\lvert G^2\rvert$	*df*	*p*-value
1	All	31	648.2				
2	Exclude LE × time	25	656.5	1	8.3	6	.22
3	Exclude age^2 × time	22	657.4	2	0.9	3	.83
4	Time linear	18	660.9	3	3.5	4	.48
5	Time quadratic	19	660.6	3	3.2	3	.36
6	Separate cov. age group	28	600.7	4	60.2	10	<.01

Note. G^2 = $-2\,l$ difference value, LE = Level of Education.

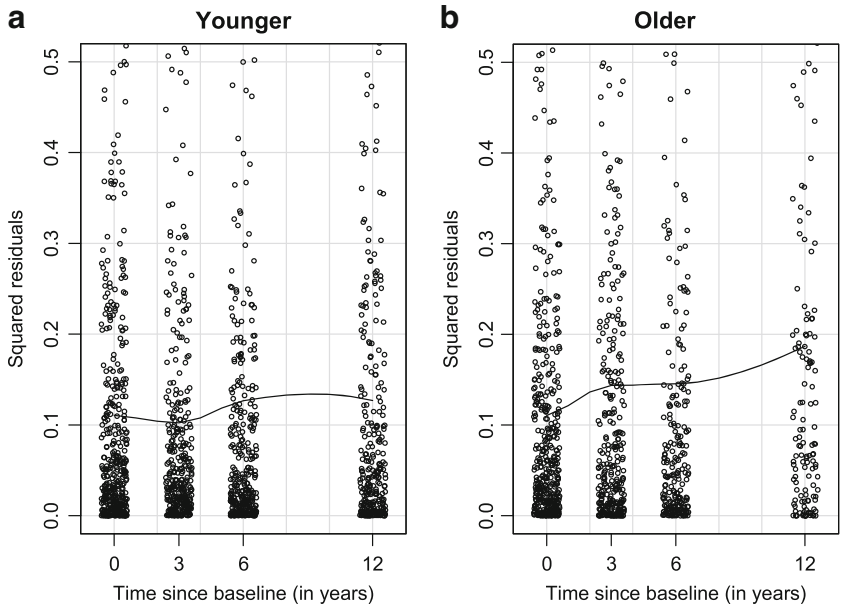

Fig. 1 Scatterplots of the squared residuals and smoothed variance functions for **a** younger participants (≤65 years at baseline) and **b** older participants (>65 years at baseline). Note that only data points that had squared residual values that were below 0.5 are presented (to depict the variance functions more clearly)

Fig. 2a for 50-, 65-, and 80-year-old females with an average LE (note that the shape of these plots is identical for males and for people with a low or a high educational level—i.e., the predicted log(SCWT) values are the same up to a constant). There was also a small (but significant) effect of age^2.

The standard linear mixed model

The preliminary mean structure of the initial standard LMM was identical to the mean structure in the initial multivariate regression model (see above). A random intercept and two random slopes (for time and time^2) were included in the preliminary covariance structure (unstructured type). We first evaluated whether the random effects were all needed in the model, by removing one random effect after the other in a hierarchical way. Note that these tests cannot be conducted by using classical likelihood ratio procedures. Instead, a mixture of two X^2 distributions should be used (with equal weights of 0.5; Verbeke & Molenberghs, 2000). The *p*-values of all the −2 *l* difference scores were significant (all *p*s < .05; data not shown), indicating that the covariance structure could not be simplified by deleting random effects from the model.

Next, the nonsignificant fixed-effect terms were removed from the model (one after the other, in a hierarchical way) to obtain a more parsimonious mean structure. This procedure yielded a model that included age, age^2, gender, LE low, LE high, time, and the age × time interaction as the covariates (see models 2 to 7 in Table 4). Finally, age-group-specific (model 8) and age group × time-specific (model 9) covariance structures were requested. The difference between models 8 and 9 is that model 8 makes the assumption that the differences in the estimated residual variances for the two age groups remain constant as a function of time (i.e., parallel variance functions for older and younger people are assumed), whilst model 9 does not make this assumption. As is shown in Table 4, model 9 had the best fit to the data.

The parameter estimates for the final standard LMM (model 9) are presented in Table 3b. In agreement with the results of the multivariate regression model, being male and having a lower LE were associated with higher log(SCWT) scores at all measurement moments. There was again a significant age × time interaction, which suggested that the increase in the log(SCWT) scores over time was more pronounced for people who were older at baseline (see Fig. 2b). The effect of age^2 was small but significant.

The linear mixed model with multiple imputation

The Markov Chain Monte Carlo method was used in the MI process to replace the missing values by 10 different imputations (Little & Rubin, 1987; Molenberghs & Verbeke, 2005; Rubin, 1996). The imputation model included the log(SCWT) scores at the different measurement moments and the covariates (i.e., age, gender, LE low, and LE high). The final standard LMM (see Table 3b) was fitted in each of the 10 "complete" data sets, and the 10 inferences were combined into a single one. The *r* statistic was computed to quantify the uncertainty portion that is stemming from incompleteness—that is, $r = \frac{(1+M^{-1})B}{\overline{\mu}}$ (where M = the number of imputations, B = the between-imputation variance, and $\overline{\mu}$ = the within-imputation variance; Schafer, 1999). The final LMM with MI model is presented in Table 3c. The age × time and time parameters had the

Table 3 The final multivariate regression model (a), standard linear mixed model (b), and linear mixed model with multiple imputation (c)

[a] Multivariate regression model					[b] Standard linear mixed model					[c] Linear mixed model with multiple imputation				
Parameter	$\widehat{\beta}_k$	*SE*	*t*	*p*	Parameter	$\widehat{\beta}_k$	*SE*	*t*	*p*	Parameter	$\widehat{\beta}_k$	*SE*	*t*	*p*
Intercept	3.880	0.025	157.75	.001	Intercept	3.879	0.025	157.19	.001	Intercept	3.878	0.018	216.62	.001
Age	0.025	0.001	18.50	.001	Age	0.025	0.001	18.47	.001	Age	0.026	0.001	24.70	.001
Age^2	0.0006	0.0002	4.08	.001	Age^2	0.0006	0.0002	4.03	.001	Age^2	0.0006	0.0001	5.09	.001
Gender	0.048	0.022	2.25	.015	Gender	0.052	0.022	2.39	.017	Gender	0.078	0.017	4.84	.001
LE low	0.178	0.024	7.57	.001	LE low	0.176	0.024	7.43	.001	LE low	0.160	0.017	9.27	.001
LE high	−0.060	0.032	−1.88	.061	LE high	−0.062	0.032	−1.93	.054	LE high	−0.080	0.025	−3.20	.002
Time	0.019	0.001	14.93	.001	Time	0.019	0.001	15.25	.001	Time	0.019	0.001	9.39	.001
Age x time	0.0012	0.0002	7.99	.001	Age x time	0.0012	0.0001	7.85	.001	Age x time	0.0011	0.0002	3.97	.001

Note. LE = Level of Education. Coding of the predictors: age = calendar age − 65; age^2 = (calendar age − 65)^2; gender, 0 = female, 1 = male; Low LE, 1 = at most primary education, 0 = otherwise; High LE, 1 = senior vocational or academic training, 0 = otherwise; time = time since baseline − 5.25.

highest *r* values (i.e., 2.28 and 1.28, respectively). The *r* values for the other covariates were substantially lower and ranged from 0.14 to 0.48. In agreement with the results of the multivariate regression model and the standard LMM, there was a significant age × time interaction, which suggested that the increase in the log(SCWT) scores over time was more pronounced for people who were older at baseline (see Fig. 2c). Being male and having a lower LE were associated with higher log(SCWT) scores at all measurement moments. The effect of age^2 was again small but significant.

Obtaining normative data

Analogously to the classical regression-based normative approach that is used in nonserial (i.e., single-measurement) testing situations (see the Introduction), three steps are needed to convert a future patient's log(SCWT) scores into percentile values. First, the expected log(SCWT) scores of patient *j* at time *t* are computed $\left(=\widehat{Y}_{tj}\right)$. Time *t* refers to the number of years since baseline. These calculations are based on the parameter estimates of the fixed effects that were provided in Table 3.

Second, the differences between the actually observed log (SCWT) scores of patient *j* at time *t* and the corresponding expected test scores are computed i.e., $\left[e_{tj}=-\left(Y_{tj}-\widehat{Y}_{tj}\right)\right]$ and standardized [i.e., $z_{tj}=e_{tj}/SD(e_{tg})$]. Note that the sign of the residuals is reversed here because a higher SCWT score is indicative of worse test performance. The $SD(e_{tg})$ values are the standard deviations of the residuals at time *t* for a person of age group *g* (younger, ≤65 years at baseline; older, >65 years at baseline) in the normative sample. These values are presented in Table 5.

Third, the standardized residuals (i.e., z_{tj}) are converted into percentile values. Histograms and QQ-plots suggested that the standardized residuals for the different models at all measurement moments were normally distributed in the MAAS (figures not shown), and Kolmogorov–Smirnov tests supported this conclusion (all *p*-values > .098). The standardized residuals can thus be converted into percentile values by means of the standard normal distribution.

An example

Suppose that a 75-year-old average educated woman with mild cognitive impairment is monitored over time. The patient was administered the SCWT at a baseline moment and 3, 6, and 12 years later. At the subsequent measurement occasions, she obtained SCWT test scores that equalled 80, 85, 90, and 100. The patient's log($SCWT_0$), log($SCWT_3$),

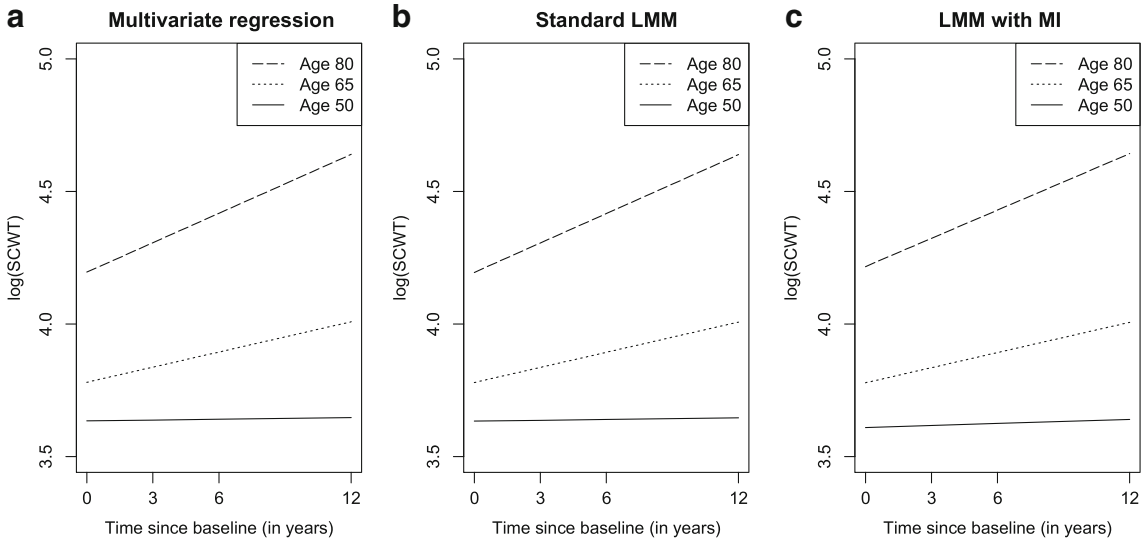

Fig. 2 Predicted marginal evolutions of the log(SCWT) values over time as based on **a** the multivariate regression model, **b** the standard LMM, and **c** the LMM with MI. The plots show the predicted values for 50-, 65- and 80-year-old females (at baseline) with an average level of education. SCWT = Stroop Color Word Test, LMM = linear mixed model, and MI = multiple imputation

log($SCWT_6$), and log($SCWT_{12}$) scores thus equalled 4.382, 4.442, 4.500, and 4.605, respectively.

The clinician uses the LMM with MI approach to evaluate the patient's test performance. This requires three steps. First, the expected log($SCWT_0$) test score is computed as based on Table 3c—that is, 4.0405 [= 3.878 + (10*0.026) + ((10^2)*0.0006) + (-5.25*0.019) + ((10*–5.25) * 0.0011)]. Second, the standardized residual is computed (as based on Table 5)—that is, −1.035 (= −(4.382 − 4.0405)/0.33). Third, the standardized residual is converted into a percentile value by means of the standard normal distribution. A standardized residual that equals −1.035 corresponds to a percentile value of 15. Thus, 15 % of the population of 75-year-old cognitively healthy females with an average level of education obtain a log($SCWT_0$) score that is equal to or higher than the score that was obtained by this woman. Using the same three-step procedure, the patient's log($SCWT_3$), log ($SCWT_6$), and log($SCWT_{12}$) scores were normed. This yielded percentile values equal to 20, 24, and 32, respectively. Thus, the SCWT test performance of the patient is within normal limits at all the measurement moments.

User-friendly normative tables

A clinician can norm the test scores of a patient by performing the required computations by hand (as was illustrated in the previous paragraph), but this procedure is time consuming and prone to making errors. To increase the user-friendliness of the normative data for clinical use, we established normative tables that present the raw SCWT scores that correspond to percentiles 5, 10, 25, 50, 75, 90, and 95, stratified by age (50, 55, . . . , 80 years), gender, and LE (the normative tables can be

Table 4 Likelihood ratio tests to evaluate the fit of a series of nested standard linear mixed models

Model	Model structure	Number of pars.	$-2l$	Ref. Model	$\lvert G^2 \rvert$	*df*	*p*-value
1	All	23	652.7				
2	Exclude age^2 × time^2	22	652.9	1	0.2	1	.65
3	Exclude age^2 × time	21	652.9	2	0	1	.00
4	Exclude age × time^2	20	655.5	3	2.6	1	.11
5	Exclude time^2 × LE	18	660.3	4	4.8	2	.09
6	Exclude time^2	17	660.7	5	0.4	1	.53
7	Exclude time × LE	15	662.9	6	2.2	2	.33
8	Separate cov. age group	16	636.8	7	26.1	1	.01
9	Separate cov. time × age	23	615.9	8	20.9	7	.01

Note. G^2 = $-2l$ difference value, LE = Level of Education.

Table 5 *SD*(*e*) values at time *t* (in years since baseline) for a person of age group *g* (younger, ≤65 years at baseline; older, >65 years at baseline), as based on the final multivariate regression model (left), the final standard linear mixed model (middle), and the final linear mixed model with multiple imputation (right)

Time *t*	Multivariate regression		Standard linear mixed model		Linear mixed model with multiple imputation	
	Younger	Older	Younger	Older	Younger	Older
0	0.33	0.33	0.33	0.33	0.33	0.33
3	0.32	0.38	0.32	0.38	0.33	0.37
6	0.36	0.38	0.36	0.38	0.37	0.39
12	0.36	0.44	0.36	0.44	0.38	0.43

downloaded at http://home.deds.nl/~wimvde/). The use of the normative tables is straightforward. For example, Table 1 in the online document immediately shows that the $SCWT_0$ score equal to 80 that was obtained by the 75-year-old average educated women of the previous example corresponds to a percentile value between 10 and 25. Note that the normative tables are based on the LMM with MI approach, because this method has some advantages over the other methods (see the Introduction and the Discussion section).

An automatic scoring program

The normative tables are easy to use but lack some accuracy, because (1) the tested patient's age has to be rounded-off if he or she is not exactly 50, 55, . . . , 80 years old and (2) because only a limited number of percentile values can be presented in the normative tables (to limit their size to a convenient format). To maximize both the user-friendliness and the accuracy of the normative data, the normative conversion procedure was implemented into an Excel worksheet (which can be downloaded at http://home.deds.nl/~wimvde/). The use of the worksheet is straightforward: the clinician simply types in the age, gender, and LE of the tested patient, together with his or her obtained raw SCWT scores at the different measurement moments, and the worksheet automatically computes the corresponding percentile values (on the basis of the LMM with MI approach).

Discussion

The multivariate regression model or the LMM?

The multivariate regression method is primarily useful when normative data have to be established for balanced data structures in which a relatively small number of repeated measurements are considered. Such data structures may arise in a longitudinal context, but they also occur in more general serial testing situations. For example, Rey's Verbal Learning Test (Rey, 1958; Van der Elst et al., 2005) is a cognitive paradigm in which a sequence of 15 words is repeatedly presented to a participant on five subsequent learning trials. Suppose that we were interested in establishing normative data for these learning trial scores. In this situation, a (almost) perfectly balanced data structure would arise in the normative sample (i.e., fixed time points are used, and missingness would be minimal), and thus the multivariate regression method would provide an adequate tool for establishing the normative data.

In situations where normative data have to be established for highly unbalanced data structures, the LMM approach is the preferred method. For example, a limitation of the present study is that the normative SCWT data can be used only to evaluate the test performance of future patients who are administered the SCWT using (approximately) the same time intervals as the ones that were used in the MAAS. In many clinical settings, variable test–retest intervals are used. For example, a first patient may be tested today, 6 months later and 5 years later, while a second patient may be tested today, 2 weeks later, and 6 weeks later (depending on the clinical profile of the patient). Suppose that we were interested in establishing a single set of normative data that allows for taking such variable test—retest intervals into account. This would, of course, require a normative sample in which variable test—retest intervals are used. For example, a study design may be used in which the lengths of the subsequent test—retest intervals are randomized per person (using some upper and lower limits of clinically relevant test–retest intervals —say, e.g., a value between 0 and 3 years). Thus, the first person in the normative sample may be tested at, for example, time points 0, 0.5, 4, 4.8, 6, 8.3, 10, and 12 years, the second person at time points 0, 2.8, 3, 4.2, 5, 6, 7.8, 9, 9.3, 10, and 11.2 years, and so on. The data structure in the normative sample would thus be highly unbalanced, but this is not a problem when the LMM is used. Only one minor modification to the normative procedure would be required. Indeed, the residual standard deviation values could no longer be computed for each time point separately but would have to be modeled as a continuous function of time (by, e.g., taking the square root of the estimated variances $\widehat{V}_i = \boldsymbol{Z}_i \widehat{\boldsymbol{D}} \boldsymbol{Z}_i' + \widehat{\sigma}^2$ Verbeke & Molenberghs, 2000).

The standard LMM or the LMM combined with MI?

The standard LMM and the LMM with MI methods are largely equivalent, provided that the imputation model includes all relationships that are considered in the analyses and the inference tasks (Molenberghs & Kenward, 2007; Molenberghs & Verbeke, 2005). Both methods have their advantages and disadvantages. The standard LMM has the advantage that it is easier to conduct and does not require a Monte Carlo component, but it has the disadvantage that it cannot handle missing covariate values. The LMM with MI, on the other hand, has the advantage that it can handle missing covariate values and missing responses simultaneously, but it has the disadvantage that it is more difficult to conduct and requires a Monte Carlo component.

In the present study, none of the covariate values were missing (because only easy-to-measure demographic covariates were considered), but there are several normative settings conceivable in which substantial missingness in the covariate values could arise. For example, suppose that one were interested in establishing IQ-corrected (rather than demographically corrected) normative data for serial test administration (see, e.g., Rentz et al., 2004). Especially in older people, it is not always straightforward to obtain IQ estimates (because of the lengthy and cognitively demanding test procedures that are typically used in IQ tests; La Rue, 1992). Missingness would thus arise in both the covariate values (i.e., the IQ scores) and the responses (i.e., the scores on the cognitive test of interest). In such situations, the use of the LMM with MI method would have the substantial advantage that it allows for using all available data in the normative analyses (while in the standard LMM, the data for people who have missing covariate values would be discarded from the analyses).

Note that the same argument applies in the context of establishing normative data for nonserial testing situations; that is, MI can be used to deal with the missing covariate values, after which classical (i.e., univariate) regression analyses can be conducted on the different completed data sets (and the inferences are combined into a single one).

Is the missingness mechanism relevant when likelihood-based methods are used?

As was noted in the Introduction, ignorable methods assume that the missingness process that generates the missing responses is MCAR or MAR (when likelihood or Bayesian inferences are chosen). In the MAAS and in most other cognitive aging studies, the MCAR assumption is not valid (Van Beijsterveldt et al., 2002). Thus, the missingness mechanism is either MAR or MNAR (Missing Not At Random; i.e., the missingness depends on unobserved data). A definitive test of MAR versus MNAR is not possible (because every MNAR model can be exactly reproduced by a MAR counterpart; Molenberghs, Beunckens, Sotto, & Kenward, 2008), but Verbeke, Molenberghs, and Rizopoulos (2010) argued that ignorable analyses provide reasonably stable results even when the MAR assumption is violated. The reason for this is that such analyses constrain the behavior of the unobserved data to be similar to the behavior of the observed data (Verbeke et al., 2010), and this is exactly what we want in the context of normative analyses. For example, suppose that a MAAS participant dropped out of the study at the second follow-up measurement occasion because he or she developed dementia. The missingness would clearly be associated with the unobserved $\log(SCWT_6)$ score (i.e., it would be MNAR), but this is not a problem, because the unknown $\log(SCWT_6)$ score of the demented patient is not of interest. Indeed, in normative studies, we are interested only in the test scores of cognitively healthy participants. When likelihood-based methods are used, the "unobserved" $\log(SCWT_6)$ and $\log(SCWT_{12})$ scores of the demented patient are modeled on the basis of the observed data of the patient at the previous measurement moments (at which the patient was still cognitively healthy) and on the basis of the observed data at all measurement moments in the normative sample. Since the observed data include only "cognitively healthy" individuals, appropriate estimates are obtained.

So, in the specific case of normative studies, the missingness mechanism is of less importance—at least, when appropriate likelihood-based methods are used. As was noted in the introduction, this is *not* the case when the regression-based change or the ANCOVA methods are used (i.e., the MCAR assumption is critical for obtaining unbiased norms when these methods are used).

No Reliable Change Indices?

Early attempts to deal with practice effects and establish norms for serial testing situations consisted of computing so-called *Reliable Change Indices* (RCIs) with correction for practice (Chelune, Naugle, Lüders, Sedlak, & Awad, 1993, see also Jacobson & Truax, 1991). The RCI method uses the *overall* mean change score and the overall *SD* (change score) in a normative sample to establish confidence intervals for change scores. By comparing the change score of a patient with these upper and lower boundaries, it can be evaluated whether the patient's performance has changed significantly (i.e., declined or improved) over time.

We did not consider the RCI method in the present study, because it is merely a special case of the regression-based change method. Indeed, when the change score is not affected by any of the demographic covariates (in the normative sample), the final regression-based change model will include only the intercept (i.e., the overall mean change

score), and the *SD*(e) value will be equal to the overall *SD* (change score). Thus, apart from the general problems that hamper the validity of the regression-based change method (see the Introduction), the RCI method has the additional limitation that it makes the (unrealistic) assumption that the change scores are not affected by any of the demographic covariates.

Some limitations of the proposed methods

The multivariate regression, LMM, and LMM with MI approaches have some substantial advantages over the regression-based change and the ANCOVA methods (see above), but these models also require some considerations that are not needed when these simple methods are used.

Some assumptions of the models

The (maximum-likelihood-based) multivariate regression and the LMM assume multivariate normality of the residuals. In addition, the LMM assumes normally distributed random effects. No straightforward procedures exist to formally test these assumptions (see Johnson & Wichern, 2007; Verbeke & Molenberghs, 2000), but this is not a severe problem. Indeed, it has been shown that the maximum likelihood estimators for the fixed effects (that are obtained under the assumption of normally distributed random effects) are consistent and asymptotically normally distributed even when the distributions of the random effects are not normal (Verbeke & Lesaffre, 1997). Similarly, the multivariate regression and LMM were shown to be robust against violations of the residual normality assumption (Jacqmin-Gadda, Sibillot, Proust, Molina, & Thiébaut, 2007).

The LMM also assumes that the variance–covariance matrix is correctly specified (to obtain unbiased estimates for the standard errors of $\widehat{\beta}$). For (sufficiently large) balanced and complete data sets, the problem of correctly specifying the covariance structure is less stringent because a robust variance estimator can be used (Liang & Zeger, 1986; Verbeke & Molenberghs, 2000). Such an estimator is consistent as long as the mean structure is correctly specified. In settings where there are missing data (which is almost always the case in serial cognitive-testing situations), a more careful reflection on the covariance structure may be warranted. Indeed, too simple a covariance structure (e.g., first-order autoregressive) could lead to bias in the mean model parameters, whereas a too complex covariance structure (e.g., unstructured) may lead to a loss of power (Molenberghs & Kenward, 2007). These issues mainly apply to situations where the sample size is small, because the loss in power that results from using an unstructured covariance matrix is negligible when the sample size is moderate to large. Since normative studies typically include moderate-to-large sample sizes (say, at least 200 people), an unstructured covariance matrix is the first choice. In normative studies with fewer participants, the specification of the covariance structure (and its impact on the established norms) should be evaluated more carefully.

Modeling time trends

The appropriateness of normative data that are established using the multivariate regression, LMM, and LMM with MI approaches depends—among other things—on the assumption that the evolution of the test scores over time is correctly modeled. Indeed, a misspecification of the model in terms of the assumed time effect (e.g., assuming a linear time trend while the true time evolution is quadratic or cubic) has an impact on the predicted test scores and the *SD*(e) values (both of which are used in the normative conversion procedure).

It is thus important to evaluate whether the assumed time effect corresponds sufficiently well to the actual time evolution. When only a small number of repeated measurements are collected, a straightforward approach is to compare the fit of a model in which time is dummy-coded (thus, a model in which no particular assumptions regarding the time evolution of the outcome are made) with the fit of a model in which a more specific time trend is assumed (e.g., a linear or a quadratic time effect). Since these models are nested, their relative fit can be formally compared by means of likelihood ratio tests (as we also did in the present study; see Table 2). When a larger number of repeated measurements are considered (which are possibly taken at different measurement occasions), it is often no longer feasible (or sensible) to dummy-code time. In this situation, the relative fit of a model in which the effect of time is captured by means of a high-degree polynomial can be compared with the fit of simpler model in which a lower-degree polynomial is used. Again, a likelihood ratio test can be used to formally compare the relative fits of the different models.

It may also be useful to take a more practical perspective and informally evaluate the extent to which the established normative data are affected by the assumed time trend (as a form of sensitivity analysis). By means of illustration, we first converted the raw test scores of the participants in the normative sample into percentile values on the basis of the final LMM that was presented in Table 3b (in which a linear time effect was assumed). Next, the participants' raw test scores were converted into percentile values on the basis of a second LMM that contained the same covariates as the first model but in which time was dummy-coded. Thus, two percentile values were obtained for each participant at each measurement moment (on the basis of models that made different assumptions regarding the time evolution). The results indicated that the maximum absolute differences between the percentile values that were obtained with both

models equalled 1, 3, 1, and 1 units (at the subsequent measurement moments). Such small differences are probably not of clinical relevance, but if these differences are deemed to be too large to be acceptable in a clinical setting, one can still retain the model in which time was dummy-coded as the final model. Thus, considerations of a statistical (e.g., the results of likelihood ratio tests), as well as a practical (e.g., differences in the established norms), nature can be taken into account in the decision process on how the time trend should be optimally modeled.

General conclusion

At present, mainly the regression-based change and the ANCOVA approaches are used to establish normative data for serial cognitive assessment. These methods have the advantage that they are based on the classical (or univariate) linear regression model (which is well-known to most behavioral researchers and straightforward to perform), but they have some major disadvantages (i.e., they can only consider the data of two measurement occasions, and they cannot deal with missing values in an appropriate way).

The multivariate regression, standard LMM, and LMM with MI approaches are not hampered by these problems. The multivariate regression model is primarily applicable when a small number of repeated measurements are taken at fixed time points. As compared with the multivariate regression model, the standard LMM and the LMM with MI approaches allow for a more adequate modeling of the covariance structure. The standard LMM and the LMM with MI are largely equivalent, because they are valid under the same assumptions and neither artificially decrease nor increase the amount of information available. The advantage of the standard LMM is that it is easier to conduct and that it does not require a Monte Carlo component. On the other hand, the LMM with MI has the advantage that it can flexibly deal with missing responses and missing covariates at the same time. When MI is used, it is important that all relationships between the covariates and responses to be studied in the scientific model of interest are included in the imputation model (to avoid "imputing under the null"; Molenberghs & Verbeke, 2005).

The different normative methods were applied to the SCWT data from the MAAS. The results showed that the log(SCWT) scores were significantly affected by age, age^2, time, gender, and LE. These covariates should thus be taken into account in the construction of the normative data. There was also a significant time × age interaction, which suggested that the increase in the log(SCWT) scores over time was more pronounced for older people (as compared with their younger counterparts). These results are in line with previous findings in the cognitive aging literature (Salthouse, 1996; Schmand et al., ; Stern, 2003; Van der Elst, 2006; Van der Elst et al., 2006d). To increase the user-friendliness of the normative SCWT data, normative tables and an automatic scoring program were provided (based on the results of the LMM with MI approach).

References

Bartels, C., Wegrzyn, M., Wiedl, A., Ackermann, V., & Ehrenreich, H. (2010). Practice effects in healthy adults: A longitudinal study on frequent repetitive cognitive testing. *BMC Neuroscience, 11,* 111–118.

Beunckens, C., Molenberghs, G., & Kenward, M. G. (2005). Direct likelihood analysis versus simple forms of imputation for missing data in randomized clinical trials. *Clinical Trials, 2,* 379–386.

Calamia, M., Markon, K., & Tranel, D. (2012). Scoring higher the second time around: Meta-analyses of practice effects in neuropsychological assessment. *The Clinical Neuropsychologist, 26,* 543–570.

Chelune, G., Naugle, R. I., Lüders, H., Sedlak, J., & Awad, I. A. (1993). Individual change after epilepsy surgery: Practice effects and base-rate information. *Neuropsychology, 7,* 41–52.

De Bie, S. E. (1987). *Standaardvragen 1987: Voorstellen voor uniformering van vraagstellingen naar achtergrondkenmerken en interviews* [Standard questions 1987: Proposal for uniformization of questions regarding background variables and interviews]. Leiden, the Netherlands: Leiden University Press.

Dikmen, S. S., Heaton, R. K., Grant, I., & Temkin, N. R. (1999). Test-retest reliability and practice effects of expanded Halstead-Reitan neuropsychological test battery. *Journal of the International Neuropsychological Society, 5,* 346–356.

Heaton, R. K., Temkin, N., Dikmen, S., Avitable, N., Taylor, M. J., Marcotte, T. D., & Grant, I. (2001). Detecting change: A comparison of three neuropsychological methods, using normal and clinical samples. *Archives of Clinical Neuropsychology, 16,* 75–91.

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59,* 12–19.

Jacqmin-Gadda, H., Sibillot, S., Proust, C., Molina, J., & Thiébaut, R. (2007). Robustness of the linear mixed model to misspecified error distribution. *Computational Statistics & Data Analysis, 51,* 5142–5154.

Johnson, R. A., & Wichern, D. W. (2007). *Applied multivariate statistical analysis* (6th ed.). New York: Pearson Education, Inc.

Jolles, J., Houx, P. J., Van Boxtel, M. P. J., & Ponds, R. W. H. M. (1995). *Maastricht Aging Study: Determinants of Cognitive Aging.* Maastricht, the Netherlands: Neuropsych Publishers.

Kenward, M. G., & Carpenter, J. (2007). Multiple imputation: current perspectives. *Statistical Methods in Medical Research, 16,* 199–218.

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied linear statistical models* (5th ed.). New York: McGraw Hill.

La Rue, A. (1992). *Aging and Neuropsychological Assessment*. New York: Plenum Press.

Lezak, M. D., Howieson, D. B., & Loring, D. W. (2004). *Neuropsychological Assessment*. New York: Oxford University Press.

Liang, K. Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrica, 73,* 13–22.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: John Wiley & Sons.

Little, R. J., & Zhang, N. (2011). Subsample ignorable likelihood for regression analysis with missing data. *Journal of the Royal Statistical Society Series C: Applied Statistics, 60,* 591–605.

McCaffrey, R. J., Duff, K., & Westervelt, H. J. (2000). *Practitioner's guide to evaluating change with neuropsychological assessment instruments*. New York: Kluwer Academic, Plenum Publishers.

Mitrushina, M. N., Boone, K. B., Razani, J., & D'Elia, L. F. (2005). *Handbook of normative data for neuropsychological assessment* (2nd ed.). New York: Oxford University Press.

Mitrushina, M., & Satz, P. (1991). Effect of repeated administration of a neuropsychological battery in the elderly. *Journal of Clinical Psychology, 47,* 790–801.

Moering, R. G., Schinka, J. A., Mortimer, J. A., & Graves, A. B. (2004). Normative data for elderly African Americans for the Stroop Color and Word Test. *Archives of Clinical Neuropsychology, 19,* 61–71.

Molenberghs, G., Beunckens, C., Sotto, C., & Kenward, M. G. (2008). Every missingness not at random model has a missingness at random counterpart with equal fit. *Journal of the Royal Statistical Society, 70,* 371–388.

Molenberghs, G., & Kenward, M. (2007). *Missing data in clinical studies*. New York: John Wiley & Sons.

Molenberghs, G., & Verbeke, G. (2005). *Models for discrete longitudinal data*. New York: Springer-Verlag.

Pasquier, F. (1999). Early diagnosis of dementia: Neuropsychology. *Journal of Neurology, 246,* 6–15.

Rapport, L. J., Brines, D. B., Axelrod, B. N., & Theisen, M. E. (1997). Full scale IQ as mediator of practice effects: The rich get richer. *The Clinical Neuropsychologist, 11,* 375–380.

Rey, A. (1958). *L'examin clinique en psychologie*. Paris, France: Presses Universitaires de France.

Rentz, D. M., Huh, T. J., Faust, R. R., Budson, A. E., Scinto, L. F. M., Sperling, R. A., & Daffner, K. R. (2004). Use of IQ-adjusted norms to predict cognitive decline in highly intelligent older individuals. *Neuropsychology, 18,* 38–49.

Rönnlund, M., & Nilsson, L. G. (2006). Adult life-span patterns in WAIS-R Block Design performance: Cross-sectional versus longitudinal age gradients and relations to demographic factors. *Intelligence, 34,* 63–78.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63,* 581–592.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association, 91,* 473–489.

Salthouse, T. A., Schroeder, D. H., & Ferrer, E. (2004). Estimating retest effects in longitudinal assessments of cognitive functioning in adults between 18 and 60 years of age. *Developmental Psychology, 40,* 813–822.

Salthouse, T. A. (1996). The processing-speed theory of adult age differences in cognition. *Psychological Review, 103,* 403–428.

Schafer, J. L. (1999). Multiple imputation: A primer. *Statistical Methods in Medical Research, 8,* 3–15.

Schmand, B., Smit, J. H., Geerlings, M. I., & Lindeboom, J. (1997). The effects of intelligence and education on the development of dementia. *A test of the brain reserve hypothesis. Psychological Medicine, 27,* 1337–1344.

Stern, Y. (2003). The concept of cognitive reserve: A catalyst for research. *Journal of Clinical and Experimental Neuropsychology, 25,* 589–593.

Strauss, E., Sherman, E. M. S., & Spreen, O. (2006). *A Compendium of neuropsychological tests: Administration, norms, and commentary* (3rd ed.). New York: Oxford University Press.

Stroop, J. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18,* 643–662.

Stuss, D., Stethem, L., & Poirier, C. (1987). Comparison of three tests of attention and rapid information processing across six age groups. *The Clinical Neuropsychologist, 1,* 139–152.

Temkin, N. R., Heaton, R. K., Grant, I., & Dikmen, S. S. (1999). Detecting significant change in neuropsychological test performance: A comparison of four models. *Journal of the International Neuropsychological Society, 5,* 357–369.

Testa, S. M., Winicki, J. M., Pearlson, G. D., Gordon, B., & Schretlen, D. J. (2009). Accounting for estimated IQ in neuropsychological test performance with regression-based techniques. *Journal of the International Neuropsychological Society, 15,* 1012–1022.

Theisen, M. E., Rapport, L. J., Axelrod, B. N., & Brines, D. B. (1998). Effects of practice in repeated administrations of the Wechsler Memory Scale-Revised in normal adults. *Assessment, 5,* 85–92.

Van Beijsterveldt, C. E. M., Van Boxtel, M. P. J., Bosma, H., Houx, P. J., Buntix, F., & Jolles, J. (2002). Predictors of attrition in a longitudinal cognitive aging study: The Maastricht Aging Study (MAAS). *Journal of Clinical Epidemiology, 55,* 216–223.

Van Breukelen, G. J. P., & Vlaeyen, J. W. S. (2005). Norming clinical questionnaires with multiple regression: The Pain Cognition List. *Psychological Assessment, 17,* 336–344.

Van der Elst, W. (2006). *The Neuropsychometrics of Aging. Normative studies in the Maastricht Aging Study*. Maastricht, The Netherlands: Neuropsy publishers.

Van der Elst, W., Van Boxtel, M. P. J., Van Breukelen, G. J. P., & Jolles, J. (2005). Rey's Verbal Learning Test: Normative data for 1,855 healthy participants aged 24-81 years and the influence of age, sex, education, and mode of presentation. *Journal of the International Neuropsychological Society, 11,* 290–302.

Van der Elst, W., Van Boxtel, M. P. J., Van Breukelen, G. J. P., & Jolles, J. (2006a). The Concept Shifting Test: Adult normative data. *Psychological Assessment, 18,* 424–432.

Van der Elst, W., Van Boxtel, M. P. J., Van Breukelen, G. J. P., & Jolles, J. (2006b). The Letter Digit Substitution Test: Normative data for 1,858 healthy participants aged 24-81 from the Maastricht Aging Study (MAAS): Influence of age, education, and sex. *Journal of Clinical and Experimental Neuropsychology, 28,* 998–1009.

Van der Elst, W., Van Boxtel, M. P. J., Van Breukelen, G. J. P., & Jolles, J. (2006c). Normative data for the Animal, Profession and Letter M naming Verbal Fluency Tests for Dutch speaking participants and the effects of age, education, and sex. *Journal of the International Neuropsychological Society, 12,* 80–89.

Van der Elst, W., Van Boxtel, M. P. J., Van Breukelen, G. J. P., & Jolles, J. (2006d). The Stroop Color-Word Test: Influence of age, sex, and education; and normative data for a large sample across the adult age range. *Assessment, 13,* 62–79.

Van der Elst, W., Van Boxtel, M. P. J., Van Breukelen, G. J. P., & Jolles, J. (2008). Detecting the significance of changes in performance on the Stroop Color-Word Test, Verbal Learning Test of Rey, and Letter Digit Substitution Test after a test-retest interval of three years: The regression-based change approach. *Journal of the International Neuropsychological Society, 14,* 71–80.

Van der Elst, W., Ouwehand, C., van Rijn, P., Lee, N., Van Boxtel, M. P. J., & Jolles, J. (in press). The shortened Raven Standard Progressive Matrices: Item Response Theory-based psychometric analyses and normative data. *Assessment.*

Verbeke, G., & Lesaffre, E. (1997). The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data. *Computational Statistics & Data Analysis, 53,* 541–556.

Verbeke, G., & Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. New York: Springer-Verlag.

Verbeke, G., Molenberghs, G., & Rizopoulos, R. (2010). Random effects models for longitudinal data. In K. van Montfort, J. Oud, & A. Satorra (Eds.), *Longitudinal Research with Latent Variables* (pp. 49–96). New York: Springer-Verlag.